CODES WITHOUT COMMAS

By F. H. C. Crick, J. S. Griffith, and L. E. Orgel

MEDICAL RESEARCH COUNCIL UNIT, CAVENDISH LABORATORY, AND DEPARTMENT OF THEORETICAL CHEMISTRY, CAMBRIDGE, ENGLAND

Communicated by G. Gamow, February 11, 1957

This paper deals with a mathematical problem which arose in connection with protein synthesis. We present the solution here because it gives the "magic number" 20, so that our answer may perhaps be of biological significance. To make this clear, we sketch in the biochemical background first.

It is assumed in one of the more popular theories of protein synthesis that amino acids are ordered on a nucleic acid strand (see, for example, Dounce[1]) and that the order of the amino acids is determined by the order of the nucleotides of the nucleic acid. There are some twenty naturally occurring amino acids commonly found in proteins, but (usually) only four different nucleotides. The problem of how a sequence of four things (nucleotides) can determine a sequence of twenty things (amino acids) is known as the "coding" problem.

This problem is a formal one. In essence, it is not concerned with either the chemical steps or the details of the stereochemistry. It is not even essential to specify whether RNA or DNA is the nucleic acid being considered. Naturally, all these points are of the greatest interest, but they are only indirectly involved in the formal problem of coding.

The first definite proposal was made by Gamow.[2] His code, which was suggested by the structure of DNA, was of the "overlapping" type. The meaning of this is illustrated in Figure 1. Gamow's code was also "degenerate"—that is, several sets of three letters (picked in a special way) stood for a particular amino acid. However, all the 64 (4 × 4 × 4) possible sets of three letters stood for one amino acid or another, so that any sequence whatever of the four letters stood for a definite sequence of amino acids.

It is easy to see that codes of the overlapping type impose severe restrictions on the allowed amino acid sequences. Unfortunately, no such restrictions have been found, although considerable (unpublished) efforts have been made, by a number of workers, to find them. Part of this work has been reviewed by Gamow, Rich, and Yčas.[3] However, the amino acid sequences so far determined experimentally are of limited extent, and it is possible that there may be restrictions on the neighbors of the rarer amino acids, such as tryptophan. Thus, while overlapping codes seem highly unlikely, partial overlapping is not impossible. At the moment, however, nonoverlapping codes seem the most probable, and these are the only ones we shall consider here.

	B	*C*	*A*	*C*	*D*	*D*	*A*	*B*	*A*	*B*	*D*	*C*
Overlapping code	*B*	*C*	*A*									
		C	*A*	*C*								
			A	*C*	*D*							
				C	*D*	*D*						
Partial overlapping code	*B*	*C*	*A*									
			A	*C*	*D*							
					D	*D*	*A*					
							A	*B*	*A*			
Nonoverlapping code	*B*	*C*	*A*									
				C	*D*	*D*						
							A	*B*	*A*			
										B	*D*	*C*

FIG. 1.—The letters *A*, *B*, *C*, and *D* stand for the four bases of the four common nucleotides. The top row of letters represents an imaginary sequence of them. In the codes illustrated here each set of three letters represents an amino acid. The diagram shows how the first four amino acids of a sequence are coded in the three classes of codes.

If each amino acid were coded by *two* bases (rather than the three shown in Fig. 1), we should only be able to code 4 × 4 = 16 amino acids. It is natural, therefore, to consider nonoverlapping codes in which *three* bases code each amino acid. This confronts us with two difficulties: (1) Since there are 4 × 4 × 4 = 64 different triplets of four nucleotides, why are there not 64 kinds of amino acids? (2) In reading the code, how does one know how to choose the groups of three? This difficulty is illustrated in Figure 2. The second difficulty could be overcome by reading off from one end of the string of letters, but for reasons we shall explain later we consider an alternative method here.

. . . , *B C A*, *C D D*, *A B A*, *B D C*, . . .

or

. . . . *B*, *C A C*, *D D A*, *B A B*, *D C* . . .

FIG. 2.—The commas divide the string of letters into groups of three, each representing one amino acid. If the ends of the string of letters are not available, this can be done in more than one way, as illustrated. The problem is how to read the code if the commas are rubbed out, i.e., a comma-less code.

We shall assume that there are certain sequences of three nucleotides with which an amino acid can be associated and certain others for which this is not possible. Using the metaphors of coding, we say that some of the 64 triplets make sense and some make nonsense. We further assume that all possible sequences of the *amino acids* may occur (that is, can be coded) and that at every point in the string of letters one can only read "sense" in the correct way. This is illustrated in Figure 3. In other words, any two triplets which make sense can be put side by side, and yet the overlapping triplets so formed must always be nonsense.

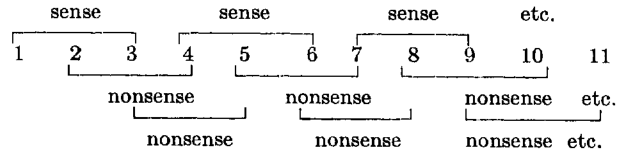

FIG. 3.—The numbers represent the positions occupied by the four letters *A*, *B*, *C*, and *D*. It is shown which triplets make sense and which nonsense.

It is obvious that with these restrictions one will be unable to code 64 different amino acids. The mathematical problem is to find the maximum number that can be coded. We shall show (1) that the maximum number cannot be greater than 20 and (2) that a solution for 20 can be given.

To prove the first point, we consider for the moment the restrictions imposed by placing each amino acid next to itself. Then, clearly, the triplet *AAA* must be nonsense, since, if it corresponded to an amino acid, α, then $\alpha\alpha$ would be *AAAAAA*, and this sequence can be misinterpreted by associating α with the second to fourth, or third to fifth, letters. We can thus reject *AAA*, *BBB*, *CCC*, and *DDD*.

It is easy to see that the 60 remaining triplets can be grouped into 20 sets of three, each set of three being cyclic permutations of one another. Consider as an example *ABC* and its cyclic permutations *BCA* and *CAB*. It is clear that we can choose any one of these, but not more than one. For suppose that we let *BCA* stand for the amino acid β; then $\beta\beta$ is *BCABCA*, and so *CAB* and *ABC* must, by our rules, be nonsense. Since we can choose at the most one triplet from each cyclic set, we cannot choose more than 20. No solution is possible, therefore, which codes more than 20 different amino acids.

We have so far not considered the effects of putting unlike amino acids together, to give pairs of the form $\alpha\beta$ and $\beta\alpha$. It might be thought that this would still further reduce the possible number of amino acids, but this turns out not to be so, since we can write down a construction which obeys all our rules and yet codes 20 different amino acids. One possible solution is

$$A\ B\ \begin{matrix}A\\B\end{matrix} \qquad \begin{matrix}A\\B\end{matrix}\ C\ \begin{matrix}A\\B\\C\end{matrix} \qquad \begin{matrix}A\\B\\C\end{matrix}\ D\ \begin{matrix}A\\B\\C\\D\end{matrix}$$

where $A\ B\ \begin{matrix}A\\B\end{matrix}$ means *ABA* and *ABB*, etc. It is easy to see, by systematic enumeration, that one can place any two triplets of this set next to each other without producing overlapping triplets which belong to the set.

The solution given above is not unique. Another satisfactory choice of 20 allowed sequences is

$$\begin{array}{ccc} B & A & A \\ & & \\ & & \\ A & B & B \end{array} \qquad \begin{array}{c} A \\ B \end{array} \; C \; \begin{array}{c} A \\ B \\ C \end{array} \qquad \begin{array}{c} A \\ B \\ C \end{array} \; D \; \begin{array}{c} A \\ B \\ C \\ D \end{array}$$

If we exclude trivial variations, such as permuting letters (e.g., A into C and C into A) or writing the code backwards, there are at least 8 different solutions. These can be obtained by taking one or the other of the two solutions given above and reversing either the entire second set of triplets, or the entire third set, or both. For example, if we reverse the second set of triplets in the first solution given above, we obtain the solution

$$A \; B \; \begin{array}{c} A \\ B \end{array} \qquad \begin{array}{c} A \\ B \\ C \end{array} \; C \; \begin{array}{c} A \\ B \end{array} \qquad \begin{array}{c} A \\ B \\ C \end{array} \; D \; \begin{array}{c} A \\ B \\ C \\ D \end{array}$$

If we enumerate *all* solutions we have been able to find, including the variations produced by interchanging letters and reversing the direction of the code, we obtain a total of 288 solutions (192 from variants of the first solution above and 96 from variants of the second one).

The problem we have considered is a special case of the more general situation in which one Greek letter is determined by n Roman letters selected from a total of m different Roman letters. One can obtain an upper limit for the number of possible Greek letters by the methods we have used, but it is not in general easy to see whether this upper limit can be achieved. One can easily see by trial that the upper limit of six, corresponding to $n = 2$, $m = 4$, cannot be achieved, only five Greek letters being possible; hence the upper limit cannot be achieved for $n = 2$, $m > 4$, either. The solution for $n = 3$ and arbitrary m is

$$A \; B \; \begin{array}{c} A \\ B \end{array} \qquad \begin{array}{c} A \\ B \end{array} \; C \; \begin{array}{c} A \\ B \\ C \end{array} \ldots\ldots\ldots\ldots\ldots \begin{array}{c} A \\ B \\ C \\ \cdot \\ \cdot \\ L \end{array} \; K \; \begin{array}{c} A \\ B \\ C \\ \cdot \\ \cdot \\ \cdot \\ L \\ M \end{array}$$

or, more concisely, writing $A_1 A_2 \ldots A_m$ for the nucleotides, then a solution which attains the upper limit is the set of triplets $A_i A_j A_k$ for all $i, j, k = 1, 2, \ldots, m$, satisfying $k \leqslant j, i < j$. We have not solved the general problem.

A Physical Interpretation.—To fix ideas, we shall describe a simple model to illustrate the advantages of such a code. Imagine that a single chain of RNA, held in a regular configuration, is the template. Let the intermediates in protein synthesis be 20 distinct molecules, each consisting of a trinucleotide chemically attached to one amino acid. The bases of each trinucleotide are chosen according to the code given above. Let these intermediate molecules combine, by hydrogen bonding between bases, with the RNA template and there await polymerization. Now imagine that such an amino acid–trinucleotide were to diffuse into an incorrect place on the template, such that *two* of its bases were hydrogen-bonded, though not the third. We postulate that this incomplete attachment will only retain the intermediate for a very brief time (for example, less than 1 millisecond) before the latter breaks loose and diffuses elsewhere. However, when it eventually diffuses to the correct place, it will be held by hydrogen bonds to all three bases and will thus be retained, on the average, for a much longer time (say, seconds or minutes). Now the code we have described insures that this more lengthy attachment can occur only at the points where the intermediate is needed. If one of the 20 intermediates could stay for a long time on one of the false positions, it would effectively block the two positions it was straddling and hold up the polymerization process. Our code makes this impossible. This scheme, therefore, allows the intermediates to accumulate at the *correct* positions on the template without ever blocking the process by settling, except momentarily, in the wrong place. It is this feature which gives it an advantage over schemes in which the intermediates are compelled to combine with the template one after the other in the correct order.

The example given here is only for illustration, but it brings out the physical idea behind the concept of a comma-less code.

In passing, it should be mentioned that while the idea of making three nonoverlapping nucleotides code for one amino acid at first sight entails certain stereochemical difficulties, these are not insuperable if it is assumed that the polypeptide chain, *when polymerized*, does not remain attached to the template. A detailed scheme along these lines has been described to us by Dr. S. Brenner (personal communication).

General Remarks.—The arguments and assumptions which we have had to employ to deduce this code are too precarious for us to feel much confidence in it on purely theoretical grounds. We put it forward because it gives the magic number—20—in a neat manner and from reasonable physical postulates. It should be noted, however, that other codes can be derived which restrict the amino acids to 20, in particular the "combination code" of Gamow and Yčas,[4] though we regard the physical assumption underlying their code as implausible. Some direct experimental support is therefore required before our idea can be regarded as anything more than a tentative hypothesis.

Summary.—The problem of how, in protein synthesis, a sequence of four things (nucleotides) determines a sequence of many more things (amino acids) is known as the coding problem. We consider codes involving nonoverlapping triplets of nucleotides, each triplet coding for one amino acid. We show that to allow all possible amino acid sequences without giving false readings of the code (due to reading the last part of one triplet and the first part of the next), we must limit the number of kinds of amino acids which the code can handle. We prove that an upper bound is 20 and show that a code for 20 can in fact be written down. It is

well known that 20 is the number found experimentally. The physical ideas behind such a code are briefly discussed.

[1] A. L. Dounce, *Enzymologia*, **15**, 251, 1952.

[2] G. Gamow, *Nature*, **173**, 318, 1954; *Kgl. Danske Videnskab. Selskab Biol. Medd.*, **22**, 3, 1954.

[3] G. Gamow, A. Rich, and M. Yčas, *Advances in Biol. and Med. Physics*, Vol. **4** (New York: Academic Press Inc., 1955).

[4] G. Gamow and M. Yčas, these PROCEEDINGS, **41**, 1011, 1955.